Systems and Computers in Japan, Vol. 17, No. 7, 1986
Translated from Denshi Tsushin Gakkai Ronbunshi, Vol. 68-D, No. 9, September 1985, pp. 1609-1616

Performance Evaluation for Memory Subsystem of Hierarchical Disk-Cache

Taizo Miyachi, Member

Institute for New Generation Computer Technology, Tokyo, Japan 108

Akitoshi Mitsuishi, Member and Tetsuo Mizoguchi, Nonmember

Information Systems and Electronics Development Laboratory, Mitsubishi Electric Corp., Kamakura, Japan 247

SUMMARY

The memory subsystem hierarchical disk-cache (MESSIAH) discussed in this paper aims at realization of high-speed access to secondary memory. MESSIAH uses a large-capacity buffer memory (A) for the input-output device and a small-capacity buffer (B) for each disk-device. (A) realizes the traditional disk-cache function. (B) realizes reduction of the disk-rotations waiting time due to a miss in RPS (rotational position sensing), reduction of overlap between disk write-in and search time, and immediate delivery of look-ahead data. In other words, in MESSIAH (A) realizes a cache memory of sufficient capacity and (B) copes with RPS misses, which increase with size of the transfer block between disk and (A). (B) also realizes high-speed write-in and read-out of the disk. Thus, MESSIAH provides greater advantage than the multiple effects of disk-cache function and the B-disk (buffer-installed disk device). This paper proposes an architecture for MESSIAH and verifies its usefulness by simulation.

1. Introduction

With recent advances in the processing function of the CPU in computer systems bottlenecks increasingly occur in input-output processing. This tendency will further be enhanced in the future. In such a situation high-speed access to the disk becomes one of the most important problems. There are various means for high-speed access to the disk, such as addition of disks and channels, modification of the blocking factors, introduction of disk-cache [2, 3], expansion of main memory, replacement of memory by high-speed devices like the CCD and magnetic bubble. Another method of the improvement is the buffer-installed disk-device (B-disk) proposed by the authors [5]. B-disk has the following three features: (1) the waiting time due to an RPS miss [1] is reduced (A); (2) in write-in, the seek operation and data transfer overlap which apparently eliminates the seek time and the rotation waiting time (B); (3) in readout, if look-ahead data exist in the data buffer, the rotation-waiting time is eliminated (C).

The authors have proposed a B-disk for high-speed input-output processing in order to match the drastic improvement in CPU processing ability. This paper further proposes a memory subsystem for hierarchical disk-cache (MESSIAH) in order to realize high-speed input-output processing combining B-disk and disk cache (DC) with the multiplying effect. Performance evaluation of DC and MESSIAH is made for the system working at an actual site.

2. Hierarchical Disk-Cache Subsystem

2.1 Reduction of input-output processing time

The disk cache (DC) is a direct method of reducing the input-output processing time. First, we explain the mechanism whereby the DC can improve the input-output processing time. The DC control scheme is also described. The DC improves the input-output processing time by virtue of the following fact, considering that the disk is a rotating medium: (1) the seek time is improved

ISSN0882-1666/86/0007-0086$7.50/0

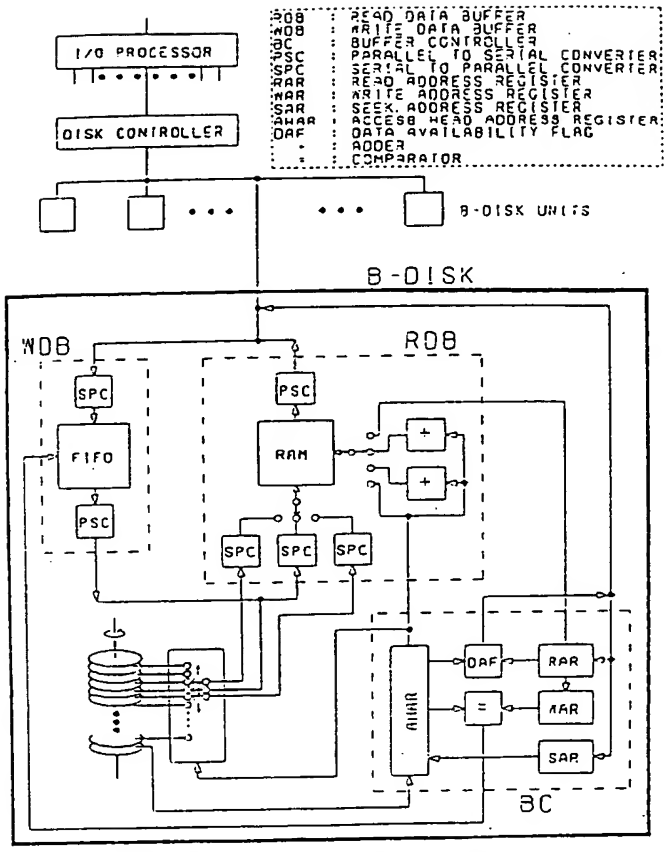

Fig. 1. Configuration of buffer-contained disk unit.

by reducing the number of arm shifts in the disk device (D); (2) the rotation waiting time is reduced by reducing the number of disk rotation waits (E); (3) on the other hand, data transfer between the disk and DC is increased, involving an increase of data transfer time (F).

Thus, in order to realize the full effectiveness of DC, the following relation should be satisfied with a sufficient margin:

$$(\text{effect of D}) + (\text{effect of E}) \gg (\text{effect of F}) \qquad (1)$$

A number of schemes has been proposed for DC control (e.g., write-through, write-after, bypass mode, sequential access mode [2] and a mode using DC only [2]. As is seen from the principle of improvement of the input-output processing time, the merit of B-disk (C) is the same as that (E) of the readout from DC. It is also seen that the input-output processing time can be improved with respect to each of the above features.

It should be noted that, even when the environment does not permit full realization of the B-disk advantages, the performance of the B-disk is not degraded as much as in other magnetic disk devices.

2.2 Structure and operation of buffer-installed disk device

B-disk is one of the important elements in MESSIAH. The B-disk structure and operation are described below. The RPS function of the disk device improves the utilization efficiency of the disk controller (DKC) and reduces the rotation waiting time. On the other hand, with an increase in busy ratio of DKC the RPS misses increase, increasing the excess rotation waiting time. B-disk is a device installed in the buffer to minimize the above waiting time.

Figure 1 shows the detailed structure of the B-disk. In this figure, RDB is the read buffer and WDB is the write buffer. The buffer controller (BC) controls RDB and WDB. RDB stores three tracks of data, which are the seek track and two adjacent tracks. It stores the data on the tracks independently of request from the input-output control device (IOC). WDB is a FIFO memory

which stores the data to be written into the disk. It receives the data from DKC without waiting for the end of the seek operation and the rotation wait. When the seek and rotation wait operations are completed, data transfer from WBD to the disk is started.

By virtue of the above functions data transfer between the disk and DKC need not be synchronized to the disk rotation, as in the past, which reduces the waiting time due to RPS misses. WDB transfers the data to be written in parallel to the seek operation during the write period, which eliminates the seek time and the rotation waiting time ((B) and (C) in Sect. 1).

2.3 Structure and operation of MESSIAH

Here we describe the memory subsystem of the hierarchical disk cache (MESSIAH) which can realize the merits of both B-disk and DC by a hierarchical arrangement of the two schemes. Figure 2 shows the structure of MESSIAH, wherein the ordinary magnetic disk is replaced by B-disk and DC is installed in the input-output processing device. The reasons for placing DC in the input-output devices are: (1) no load is further placed on CPU and (2) as many B-disks as possible can share DC, increasing effectively the memory capacity of DC.

As is seen in the structure of MESSIAH high-speed input-output processing is realized by adding a small-capacity buffer memory to the magnetic disk device and a large-capacity buffer memory to the input-output device of the ordinary computer system.

The reason for the high-speed input-output processing of MESSIAH is, in addition to the simultaneous realization of the merits of B-disk and DC, to the remedy of the DC shortcoming described in (F). In general, in order to improve the hit ratio of DC, as much data as possible should be read ahead. This, however, increases the data transfer and size of the transfer block between B-disk and DC, leading to more frequent RPS misses. On the other hand, B-disk can retain the object data in the buffer memory, which eliminates the need for the disk-rotation waiting in the event of RPS miss. For this reason, B-disk can prevent performance degradation due to frequent RPS misses, which has been a serious problem in DC. In this sense, MESSIAH realizes the multiplying effects of B-disk and DC.

Thus, the merits of MESSIAH can be summarized in the following six points.

(1) The seek time is reduced by reducing the number of arm shifts in the disk device (B-disk).

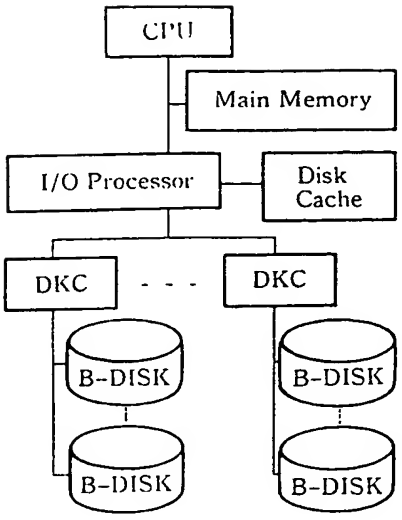

Fig. 2. Configuration of MESSIAH.

(2) The rotation waiting time is reduced by reducing the number of disk-rotation waits.

(3) The waiting time in RPS misses is reduced.

(4) In writing, the seek operation and data transfer overlaps, which eliminates the seek time and the rotation waiting time for the disk device (B-disk).

(5) In reading, if the requested data exist in the B-disk buffer, the rotation waiting time is eliminated.

(6) The system can cope with frequent RPS misses due to increased block size in the data transfer between B-disk and DC.

3. Method of Performance Evaluation

Performance evaluation was made for the computer system in actual operation (called DB 16 in this paper). DC 16 is a system with 300-MB disk, 16 spindles with optimal file location (system disk: 2 spindles and private disk: 14 spindles). In DB 16 the database processing and file-batch processing jobs are executed in 8-tuple jobs. A CODASYL type database is used.

3.1 Flow of performance evaluation

The performance was evaluated by the following procedure (Fig. 3).

(1) From statistical data of the actual system (e.g., I/O frequency) the analysis period is selected. The disk access pattern of the trace data is analyzed. The

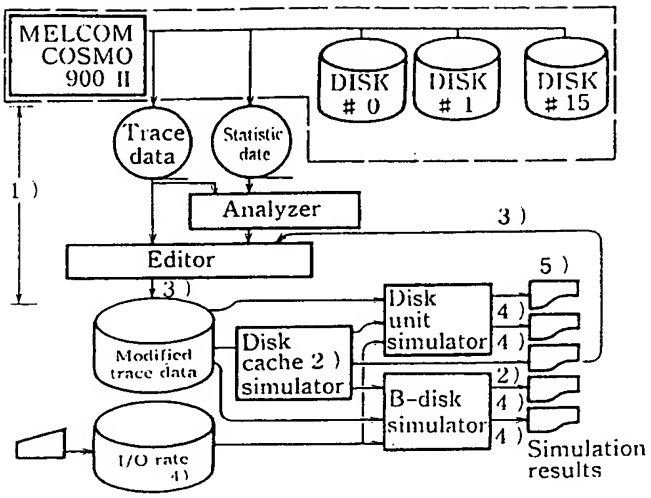

Fig. 3. Procedure of performance evaluation for MESSIAH.

The trace data are extracted and edited for DC performance evaluation.

(2) Performance of DC is evaluated.

(3) From the result of performance evaluation of DC, the trace data are selected for performance evaluation of MESSIAH and DC considering the RPS misses.

(4) The trace data are edited for performance evaluation of MESSIAH and DC considering the RPS misses and the request rate of I/O is calculated.

(5) The performance evaluation is made for MESSIAH and DC considering RPS misses.

3.2 Simulator for disk cache

The DC simulator uses the trace data of the disk access and evaluates the DC performance by calculating the cache hit ratio, improvement of access time and hit-depth pattern. The access time is calculated considering the seek length of the disk device. In the simulator, DC is placed in the input-output processing device.

A feature of the simulator is that 4 kinds of block sizes (1 page = 2 KB, 1 track, 2 tracks and 1 cylinder) can be selected in the data transfer between DC and magnetic disk. The unit of data management in DC is the same as the transfer-block size, in principle. The exception is the case where the transfer-block size is 2 tracks, where the unit of management is 1 track. By varying the block size of the data transfer the effect of prefetch of the physical address as affected by the disk address can be examined.

The control scheme for DC employs the write-through scheme, which possesses high reliability in file updating. The write-through scheme employed in this system is described below.

Read: (i) If the object data are in DC (hit), the data are transferred from DC to main memory).

(ii) If the object data are not in DC, the data are transferred from the disk to both DC and main memory.

Write: (i) When hit, the corresponding record and disk are updated.

(ii) When not hit, only the disk is updated.

Cache replacement is performed by LRU scheme. Since the I/O request in the object system was less than 2 KB, it is defined that hit occurs when the page requested by I/O exists in DC. The input-output service time (T_c) is defined as follows.

When hit: T_c = mean DC access time

When not hit: T_c = (seek time) + (mean rotation waiting time) + (data transfer time)

3.3 Simulation of MESSIAH

MESSIAH is simulated using a DC simulator, B-disk simulator and an I/O request edit program which simulates DC operation to produce the request from the input-output processing to B-disk.

3.4 Simulator for buffer-installed disk device

The B-disk simulator is composed of two simulators, with and without buffer. It can evaluate the B-disk performance under various traffic conditions. As the simulator for the disk without buffer, Mutsubishi M2838-F with 300-MB disk, is used as the model. The B-disk simulator employs M2838-F plus a buffer. Each of the simulators operates assuming a system composed of a CPU, a DKC and 8 disks. The IOP (I/O processor) has little effect on the performance and is not included in the model.

4. Result of Performance Evaluation

4.1 Performance evaluation of disk cache

The DC performance evaluation is made for the following 8 access patterns.

(1) Database access (A, B, L, 2, 4)

(2) Database and file access (C, T)

(3) File access (D)

As the database (DB) a general CODASYL-type DB management system is considered, which realizes high-speed retrieval using a hash function.

The data used in (1) ~ (3) were selected from large-scale typical or general data. B is an access pattern almost equal to a sequential search; A, L, 2 and 4 are random-access patterns often observed in database access. For example, in the access pattern of A the maximum and mean seek length are 692 and 49.2, respectively. C and T are access patterns in which the access is made to the database and to the file nearly the same number of times. D is a sequential access pattern.

Evaluation of cache hit ratio

Figures 4 ~ 7 show the DC hit ratio when the transfer block sizes are 1 page, 1 track, 2 track and 1 cylinder, respectively. The cache size is represented on a per-spindle basis. The maximum cache size is set as 12-MB/spindle based on the following analysis. In DC the probability that a hit is produced below depth 30 of the LRU stack is above 81% (88% in most cases). When the transfer block size is 1 cylinder (maximum), cache size of 322 KB × 30 = 9.66 MB is required in order to realize an LRU stack of depth 30.

The following properties are observed concerning the hit ratios of Figs. 4 ~ 7.

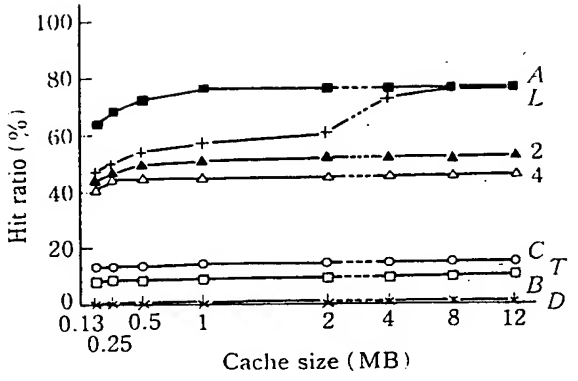

Fig. 4. Hit ratio (transfer block size = 1 page).

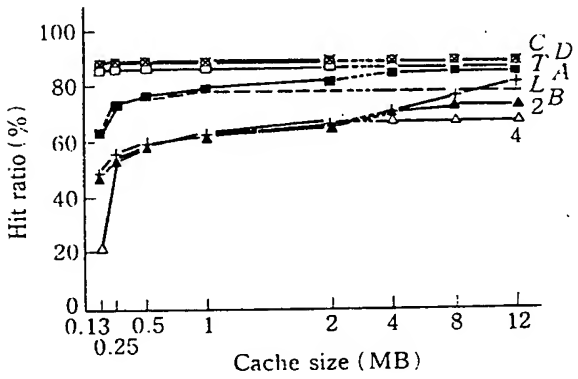

Fig. 5. Hit ratio (transfer block size = 1 track).

(1) The hit ratio saturates in most cases at cache size of 1-MB/spindle.

(2) The hit ratio increases with the transfer block size in the order of 2 tracks, 1 track and 1 page, indicating that prefetch of a neighboring disk address is effective. When the transfer block size is 1 cylinder, however, 1 ~ 2-MB/spindle is not sufficient for database access, resulting in a poor hit ratio.

Input-output service time (T_c)

Table 1 shows the input-output service time (T_c) utilizing DC for certain boundary values. The following observations are made concerning T_c for access patterns, A, B, C and D. The input-output processing speed is improved by the following features: A : 1.61 times, B : 1.43 times, C : 204 times, D : 2.38 times. The transfer block sizes are A : 1 page; B : 2 tracks; C : 2 tracks; D : 1 cylinder. For each access pattern the following

Table 1. Reduction of input-output service time (T_c) by DC

Access pattern	Transfer block size			
	1 page	1 track	2 tracks	1 cylinder
D	×	○	◎	◎
C	×	○	◎	○
T	×	○	○	△
B	×	△	△	△
A	△	△	△	×
2	△	△	△	×
L	○	△	△	×
4	○	△	△	×

(Note) T_c is reduced to less than 0.5 times in ◎, 0.5 to 0.6 times in ○, 0.6 to 0.8 times in △, and is above 0.8 times in ×.

properties are seen. In A (DB), T_c is reduced to 0.6 ~ 0.7 times, except for the case of the transfer block size of 1 cylinder. In B (DB), C (DB and file) and D (file), T_c is reduced to 0.7 ~ 0.8, 0.5 ~ 0.6 and 0.4 ~ 0.6 times, respectively.

It is seen from the above result that T_c is reduced by 0.1 ~ 0.3 more in file access than in data access. Concerning the database access, DC does not work effectively when the transfer block size is large (1 cylinder) in A and small (1 page) in B. The reason for this is that the access pattern of A is random and that of B is partly sequential. Thus, it is concluded that the access pattern is important when DC is employed in the database access.

Comparing the cases with transfer block size of 1 track and 2 tracks, the latter is seen to reduce the access time by 1 to 12%. This indicates that data fetch by the disk address is somewhat effective. The effect of prefetch is the most remarkable in D. The reason for that is that D has an almost sequential access pattern.

Input-output service time (T_r) considering RPS misses.

T_c performance evaluation discussed up to this point does not take into consideration the RPS misses produced by racing in the disk controller (DKC) by the magnetic disk devices. Consequently, as the next step a simulation was performed considering RPS misses. Table 2 shows the input-output service time for access patterns A and D as obtained by simulation and give the results

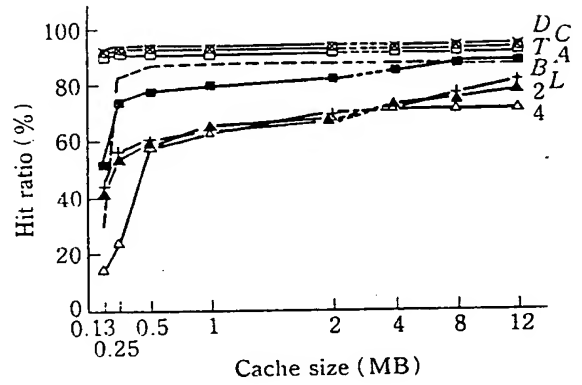

Fig. 6. Hit ratio (transfer block size = 2 tracks).

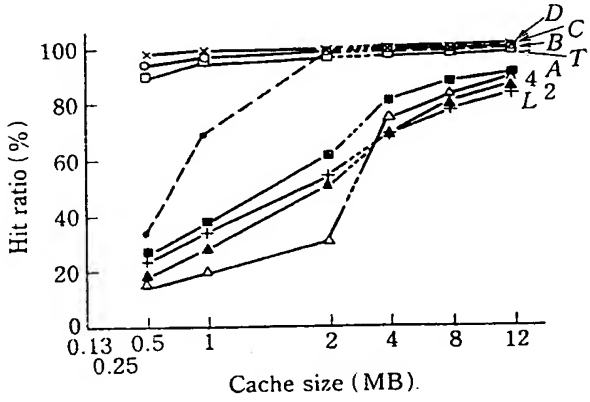

Fig. 7. Hit ratio (transfer block size = 1 cylinder).

Table 2. Input service time (T_r) considering RPS misses

Processing	DC	DC size (MB)	BS	RPS miss ratio (%)	DKC busy ratio (%)	Mean T_r (ms)
A	Without	———	max.1 P	22.1 ~ 24.1	25.4	27.1
D	Without	———	max.1 P	28.7 ~ 30.9	25.8	29.5
A	With	0.25	1 P	10.3 ~ 13.9	15.0	15.9
		12.00		11.0 ~ 12.3	13.8	14.7
		0.25	1 T	38.4 ~ 39.7	47.7	23.0
		12.00		29.7 ~ 31.2	35.0	18.1
		0.25	2 T	62.6 ~ 66.1	71.4	35.9
		12.00		47.2 ~ 51.8	52.3	23.6
D	With	1.00	1 P	28.7 ~ 30.9	25.8	29.5
		1.00	1 T	24.2 ~ 26.3	26.7	15.3
		1.00	2 T	25.3 ~ 27.0	27.2	14.3
		1.00	1 C	33.6 ~ 34.9	29.2	14.6

(Note 1) BS transfer block size, P: page, T: track, C: cylinder.
(Note 2) Processing implies the access pattern.

Table 3. Reduction ratio of T_r by MESSIAH

Processing	Use of MESSIAH	DC size (MB)	BS	Mean T_r (ms)	Mean ratio of T_r
A	Without	———	max.1 P	27.1	1.0
D	Without	———	max.1 P	29.5	1.0
A	With	0.25	1 P	9.0	0.33
		12.00		7.5	0.28
		0.25	1 T	12.5	0.46
		12.00		9.1	0.34
		0.25	2 T	25.6	0.94
		12.00		13.2	0.49
D	With	1.00	1 P	13.4	0.45
		1.00	1 T	6.5	0.22
		1.00	2 T	5.8	0.20
		1.00	1 C	2.9	0.098

(Note 1) BS: transfer block size, P: page, T: Track, C: cylinder.
(Note 2) Processing indicates the access pattern.

for (T_r = T_c + (RPS miss handling time)), DKC busy ratio and RPS miss.

It is seen from these results that DKC is a bottleneck of the system in A for transfer block size of 2 tracks. The reason for this is as follows. The threshold values for RPS misses and DKC busy ratio when DKC is the bottleneck of the system are both 30 - 40%. The situation in the simulation actually exceeds this value. The same situation is anticipated for transfer block size of 2 tracks. Thus, in the stand-alone application of DC in DB processing, the designer should be careful that DKC is not a bottleneck by referring to this result.

Table 3 shows the reduction ratio of the input-output service time (Tr) with introduction of MESSIAH. A is the random access pattern often observed in DB access and D is a sequential access pattern. It is seen that by using MESSIAH Tr can be reduced to 0.94 ~ 0.28 and 0.45 ~ 0.098 times in A and D, respectively. In terms of processing speed, the factors are 1.1 ~ 3.6 and 2.2 ~ 10.2 times in A and D, respectively.

It is seen from Table 2 and B-disk simulation that Tr can be reduced by using DC by 1.3 ~ 0.54 and 1.0 ~ 0.48 times in A and D, respectively. By using B-disk Tr can be reduced to 0.62 and 0.45 times in A and D, respectively. As a result of evaluation it is noted that the following relation is derived:

(speed improvement by MESSIAH)
> (speed improvement by B-disk)
× (speed improvement by DC) (2)

Equation (2) indicates the multiplying effect achieved by the hierarchically combined functions of B-disk and DC. This advantage seems to be due to reduction of the RPS miss time (see 2.3) upon increased transfer block size (or transfer data) between DC and magnetic disk device (B-disk).

The following merits were also verified by simulation. The write time in the disk is reduced by using B-disk. The read time is reduced for the previously accessed track and adjacent tracks. The read time is reduced by the large-capacity DC, which has been difficult to achieve with the ordinary small-capacity B-disk. The input-output processing speed is improved by decreasing the number of accesses to B-disk. From these results it is concluded that the MESSIAH structure is very useful. The size of buffer memory needed in MESSIAH is several tens KB/spindle in B-disk, and approximately 1-MB/spindle in DC.

Further speed improvement by MESSIAH

This paper evaluated the performance of a scheme whereby DC in MESSIAH is controlled by a write-through scheme, resulting in observation of the following two points.

(1) For access patterns such as A, L and 2, which are often observed in DB access, the WRITE instruction, which accesses the partial write-in of data transferred to DC by READ instruction, occupies 80 ~ 89% of the whole WRITE instruction.

(2) For access patterns such as D, which makes almost sequential access to the file, the probability that WRITE instruction for B-disk accesses a track accessed by the previous READ or WRITE instruction is 57.4 ~ 67.7% of the total READ and WRITE instructions. The probability that the same track or an adjacent track is accessed is 81.7 ~ 97.4% of the total READ and WRITE instructions.

The following observations are made from the above two points. It is seen from (1) that further improvement of the input-output processing speed can be made for such access patterns as A, which is often seen in DB access, by controlling the MESSIAH DC by a write-after scheme. It is seen from (2) that further improvement of the speed can be made for such access patterns as D, which makes an almost sequential access to the file by controlling the MESSIAH DC by a write-after scheme.

Thus, further improvement of the input-output processing speed can be made by controlling the MESSIAH DC by a write-after scheme.

5. Conclusion

This paper proposed a memory subsystem of hierarchical disk-cache (MESSIAH) which is a hierarchical combination of the functions of B-disk and DC. Performance evaluation was made by simulation for typical or general data selected from the trace data of database and file processing obtained from an actual operating system. Performance evaluations of B-disk and DC were made in order to examine the performance of MESSIAH.

As a result, it was seen that by introducing MESSIAH with write-through control of DC the speed of input-output processing can be improved up to 3.6 ~ 10.2 times in database processing (A) and file processing (D), respectively. The relation

(effect of MESSIAH) > (effect of B-disk) × (effect of DC)

was established, thus confirming by simulation that MESSIAH realizes the multiplying effect of B-disk and DC.

The size of buffer memory needed in MESSIAH in order to improve the input-output processing speed is several tens KB/spindle for B-disk and approximately 1-MB/spindle for DC, which are sufficiently practical values. The performance of MESSIAH was evaluated in this paper with the DC controlled by a write-through scheme. Further possibility of speed movement for input-output processing was investigated by adopting a write-after control scheme for DC. The detailed evaluation of this scheme is left for further study.

Acknowledgement. The authors wish to thank Dr. M. Sudo, Director, Information Electronics Laboratory, R & D Div., for providing the opportunity for this research, as well as by Mr. T. Kan, Information Electronics Laboratory, for valuable advice.

REFERENCES

1. D.T. Brown et al. Channel and direct access device architecture, IBM Syst. J., 23, 11, pp. 186-199 (1972).
2. T. Tokunaga et al. Integrated disk cache system with file adaptive control, Proc. I.E.E.E. COMPCON Fall, pp. 412-416 (1980).
3. T. Kan et al. Measurement of efficiency of disk-cache device by simulation, Trans. Inf. Proc. Soc. Jap., 22, 1, pp. 22-28 (1981).
4. T. Miyachi et al. Hierarchical disk-cache subsystem and its performance evaluation, Tech. Rep., Inf. Proc. Soc. Jap., Contr. & Evaluation of Computer Systems, 16 (1982).
5. A. Mitsuishi, T. Miyachi and T. Mizoguchi. Performance evaluation of buffer-installed disk-device, Trans. I.E.C.E., Japan, J67-D, 11, pp. 1301-1308 (Nov. 1984).
6. A. Mitsuishi et al. Buffer-installed disk-device and its performance evaluation, Proc. 24th Nat. Conv. Inf. Proc. Soc. Jap., 6D-1 (1982).
7. Y. Bard. A model of shared DASD and multipathing, C. ACM, 23, 10, pp. 564-572 (1980).
8. A.J. Smith. Optimization of IO System by Cache Disks and File Migration: A Summary, Performance Evaluation 1. North Holland Publishing Company, pp. 249-269 (1981).
9. H.J. Wojtkowiak. Effect of a Reduced Memo Residence Set on System Performance, IBM Research Report (Dec. 1980).
10. D.M. Stein. The Memory Organization of Future Large Processors, IBM Research Report (June 1979).

AUTHORS (from left to right)



Taizo Miyachi. Graduated 1977, Dept. Inf. Eng., Fac. of Eng., Shizuoka Univ. Completed Master's course 1979, Grad. Sch., Tohoku Univ. Since 1979 affiliated with Mitsubishi Electric Co. Engaged in research in database and hierarchical memory systems. Since 1982 associated with the Institute of New Generation Computer Technology and engaged in R&D of knowledge-base systems. Member Inf. Proc. Soc. Japan.

Akitoshi Mitsuishi. Graduated 1978, Dept. Appl. Physics, Fac. of Eng. Osaka Univ. Completed Master's course 1980, Grad. Sch. Osaka Univ. Since 1980 affiliated with Mitsubishi Electric Co. Engaged in R&D in performance evaluation of computer systems, hierarchical memory systems and dedicated processors. Presently with Information Electronics Laboratory, Member Inf. Proc. Soc. Jap.

Testuo Mizoguchi. Graduated Dept. Electrical Eng. 1965, Fac. of Eng. Chiba Univ. Completed 1971 course, Dept. Computer Science, Univ. Calif. Berkeley. Presently affiliated with Mitsubishi Electric Co., Inf. Electronics Lab. Engaged in research on database and office automation. Doctor of Engineering.